Modeling prediction error improves power of transcriptome-wide association studies

Kunal Bhutani[1,2,*], Abhishek Sarkar[3,4,*], Yongjin Park[3,4],
Manolis Kellis[3,4,#], Nicholas J. Schork[1,2,#]

[1] University of California, San Diego, La Jolla, CA

[2] J. Craig Venter Institute, La Jolla, CA

[3] Massachusetts Institute of Technology, Cambridge, MA

[4] Broad Institute of Harvard and MIT, Cambridge, MA

[*] Equal contribution

[#] Corresponding authors

1 Abstract

Transcriptome-wide association studies (TWAS) test for associations between imputed gene expression levels and phenotypes in GWAS cohorts using models of transcriptional regulation learned from reference transcriptomes. However, current methods for TWAS only use point estimates of imputed expression and ignore uncertainty in the prediction. We develop a novel two-stage Bayesian regression method which incorporates uncertainty in imputed gene expression and achieves higher power to detect TWAS genes than existing TWAS methods as well as standard methods based on missing value and measurement error theory. We apply our method to GTEx whole blood transcriptomes and GWAS cohorts for seven diseases from the Wellcome Trust Case Control Consortium and find 45 TWAS genes, of which 17 do not overlap previously reported case-control GWAS or differential expression associations. Surprisingly, we replicate only 2 of 40 previously reported TWAS genes after accounting for uncertainty in the prediction.

2 Introduction

Thousands of loci associated with hundreds of complex diseases have been reported in the NHGRI catalog of genome-wide association studies [25, 50] (GWASs). However, most genome-wide significant loci are devoid of protein-coding alterations [18] and likely instead affect gene regulation by mechanisms such as transcriptional, post-transcriptional, or epigenetic regulation. Several studies

have directly investigated the role of transcriptional regulation on complex diseases by jointly considering genotypes, expression, and phenotypes using Mendelian randomization [44, 3]. However, such studies require genetic, transcriptomic, and phenotypic data to be measured in all samples, which is still prohibitive at the scale of GWAS.

Recent large-scale efforts such as the Gene-Tissue Expression Project (GTEx) have generated reference transcriptomes across multiple human tissues[4]. These data have enabled transcriptome-wide association studies (TWAS), which use the reference expression data to build models of transcription regulation, impute gene expression into GWAS cohorts where expression is not measured, and directly test for association between predicted gene expression and phenotype[15, 16]. However, current methods are limited to using only point estimates of imputed expression, while ignoring the uncertainty in the predicted expression.

The impact of not incorporating uncertainty of imputed predictors on genetic association analysis has been previously studied in the context of imputed genotypes in GWAS. Although taking the best-guess genotype (posterior mode) is standard practice for GWAS, using posterior mean dosages increases power to detect associations [56, 1]. Standard methods from missing data theory such as multiple imputation have been applied in this setting yielding reductions in bias [42, 35]. More sophisticated (and computationally expensive) methods such as SNPTEST [33] can analytically integrate over the full posterior distribution of the imputed genotypes, further improving power.

Here, we develop a novel Bayesian method for modeling uncertainty in imputed expression and propagating this uncertainty through TWAS. We compare our method to existing methods for TWAS and standard methods from missing data and measurement error theory and show that our method increases power to detect genes associated with phenotype. We apply our methods to GWAS for seven diseases from the Wellcome Trust Case Control Consortium [9] and find 45 TWAS genes, replicating only 2 of 40 previously reported TWAS genes. We find 17 of the 45 genes have not yet been identified by GWAS or differential gene expression in case-control cohorts.

3 Results

3.1 Uncertainty in TWAS

The key insight enabling TWAS is that one can train models to predict gene expression from genotype in reference cohorts, and use these models to impute unobserved gene expression values in GWAS samples using the genotype information. Direct tests for association between gene expression and phenotype can be pursued, more directly identifying putative causal genes for the phenotype of interest. However, current methods only use a point estimate of the predicted gene expression in TWAS and ignore the uncertainty in the prediction. The uncertainty arises from two sources: not learning the correct model for transcriptional regulation (e.g., omitting trans-regulatory effects), which we do not consider here, and not correctly estimating the model parameters due to sample size, linkage disequilibrium, or biological and technical confounders.

Our main contribution is a novel two-stage Bayesian regression model which incorporates uncertainty for each SNP effect (BAY-TS). The key idea of BAY-TS is to use the posterior distribution of SNP effects from the first-stage regression of expression against genotype as the prior distribution

on effects in the second stage regression of phenotype against expression (Methods). After performing the second stage regression, we compute a Bayes factor comparing the fitted model against a null model where gene expression has no effect on phenotype.

We compared our method against existing methods, which merely calculate association statistics using ordinary least squares. We compared two strategies: using the full elastic net model trained on the entire GTEx dataset (OLS-E), and using the mean value of imputed expression across 50 bootstrapped models (OLS-M). We note that existing methods for TWAS implement OLS-E only.

We then compared our method to multiple imputation (MI), a standard method for handling completely missing (i.e., unobserved) data such as gene expression in TWAS [30]. For each gene, we imputed 50 expression levels for each individual using the bootstrapped models described above and estimated 50 effect sizes. We then combined these effect sizes into a single association statistic for evaluation incorporating both the mean and the variance of the 50 estimates (Methods).

Finally, we compared our method to regression calibration (RC), a standard method from measurement error theory. Measurement error theory explicitly models the error in observations (here, imputed expression) and predicts the impact of not including the errors on statistical inference [14]. Briefly, not explicitly including error in the model leads to a violation of the model assumptions and therefore leads to bias in the estimated regression coefficients. Applying this theory to TWAS, we modeled each imputed expression value as the true expression value plus additive error. We estimated the distribution of the error as the variance in predicted expression across the 50 bootstrapped models. We then performed RC, estimating the true expression based on the estimated measurement (imputation) errors and regressing phenotype on the true expression.

3.2   Simulation study

We used real genotype data to jointly simulate gene expression at both causal and non-causal genes in simulated reference and GWAS cohorts and continuous phenotypes in the GWAS cohorts, as done in prior work [16]. To calculate the recall and the effective false discovery rate (FDR), we tested single-gene associations against a phenotype generated using all of the simulated genes (Methods). For each simulated data set, we computed the the area under the precision-recall curve (AUPRC) of each method. We compared the AUPRC rather than the area under the receiver operating characteristic (AUROC) curve because the AUROC is not appropriate when the proportion of positive and negative examples is not 0.5 [?]. We ranked the genes according to the association statistic computed by each method, then computed the cumulative precision and recall (based on the simulated ground truth) for each position in the ranked list (Figure S2).

We simulated a reference cohort of 300 individuals and a GWAS cohort of 5,000 individuals for which 40 genes were causal and 1,000 were non-causal, and found that Bay-TS outperformed all other methods across the entire range of cis-regulatory architectures (Figure 1a). Interestingly, neither MI nor RC improved performance over OLS-E, which is likely explained by the fact that in this setting a central assumption in measurement error theory is violated. Specifically, when modeling the error in imputation, the second-stage predictors (means of imputed expression) and their associated errors (variances of imputed expression) are correlated because they both depend on genotype.

We then investigated the recall of each method at FDR 10%. To control the FDR of BAY-TS, we

3

calibrated a threshold for the Bayes factor (BF > 24) which controlled the average FDR over the entire range of simulation parameters at the desired level (Methods). For the other methods, we used the Benjamini-Hochberg procedure to control the FDR. We found that BAY-TS again consistently outperformed all the methods, with average recall equal to 15% (Figure 1b). Surprisingly, MI had the worst performance, likely due to deflated association statistics (Figure S3).

3.3 Application to seven diseases

Before applying our method to real data, we sought to evaluate the impact of expression normalization on the trained first-stage models in TWAS. We first asked which genes had gene expression reliably predictable from cis-genotypes. We used whole blood expression in 338 individuals from the GTEx project [4] and compared models trained on the published normalized expression values (GTEx-Norm) to models trained on regularized log transformed expression (RLog) (Methods). We found 1,666 and 1,655 genes with $R^2 \geq 0$ for GTEx-Norm and RLog, respectively, and found 1,043 genes common between the two sets (Figure 2a). However, only 987 of the 1,043 genes had enough non-missing data in the GWAS cohorts to successfully impute expression into GWAS.

We next considered the ratio between the variance of within-individual imputed expression estimates to across-individual estimates, which we define as the variance ratio (VR) (Methods). Measurement error theory predicts VR is correlated with power to detect associations, which we confirmed in simulation (Figure S4). Intuitively, after estimating the error variance of imputed expression, we can denoise the imputed expression by subtracting variation due to error in the imputed values. For the set of 987 genes, RLog based models had higher VR than GTEx-Norm based models (Figure 2b, S5), suggesting RLog normalization increases power to detect TWAS genes.

We then asked whether the same cis-SNPs were reliably selected in the first stage model fitting for the 987 genes. We note that in this setting first-stage regression predictors are correlated due to linkage disequilibrium, and therefore regularized regression techniques such as elastic net will in general not select the same non-zero regression coefficients for replicate data sets. We calculated the discordance fraction between the predictors selected in our first-stage bootstrap training against the predictors included in a single model trained on the entire dataset (Methods). We found not only that GTEx-Norm models have more predictors per gene on average, but also have higher discordance in the selected predictors. Since RLog models find a more consistent set of cis-SNPs and have larger VR (Figure S5), we perform TWAS based on RLog models.

We performed TWAS on seven disease cohorts from the Wellcome Trust Case Control Consortium [9]: bipolar disorder (BD), Crohn's disease (CD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), Type 1 Diabetes (T1D), and Type 2 Diabetes (T1D). We first sought to replicate PrediXcan by using the published cis-regulatory model weights on our imputed genotypes [15]. We replicated only 18 of 40 reported TWAS genes (Bonferroni correction, $p < 5.76 \times 10^{-6}$), likely due to differences in imputation pipelines between the two studies. 13 of the discordant genes are in the Major Histocompatability Complex (MHC) region, and 15 have estimated effect sizes with the same sign as previously reported (Table S1).

We then performed TWAS using BAY-TS in each of the seven diseases and found 45 associated genes (FDR 10%, Table 1). Surprisingly, we only replicated 2 of 40 reported TWAS genes (Table S1). Moreover, we found that only 9 of the 40 reported genes in our set of high confidence genes

with reliably predictable expression. The discrepancy in TWAS genes is due partly to differences in genotype imputation, expression normalization, and first stage model fitting as described above, and highlight the importance of data processing choices when performing TWAS. We compared BAY-TS to the other methods proposed above (including OLS-E, equivalent to PrediXcan [15]) and found that 23 of the 45 genes are not found by any of the other methods (Table S2, S3, S4). The other methods collectively did not find any associations for BD or CAD, likely due to lack of power.

We discovered three genes significantly associated with BD: SARDH, ZNF79, and WDR25. A higher burden of CNVs in the SARDH gene has been previously linked to BD and Schizophrenia[27], and its loss is the principal cause for sarcosinemia, which often leads to mental impairment[41]. However, ZNF79 and WDR25 are both poorly characterized and have not been studied in the context of BD previously.

We discovered four genes associated with CAD: DAGLB, CYTH3, PARL, and TMEM158. DAGLB is active in the triglyceride lipase activity pathway, and has been linked previously to high-density lipoprotein levels levels[51]. CYTH3 is a regulator of PI-3 kinase signaling, which mediates many pathways in the cardiovascular system [43]. Variants in PARL have been associated with increased levels of plasma insulin and predisposition to CAD [37]. In Chinese populations, the gene was also linked to higher levels of triglyceride and total cholestrol in both T2D and control populations [31]. Lower TMEM158 expression is associated with increased risk in both of the CAD and T2D populations in our study. It has previously transcriptionally associated with T1D, T2D, and gestational diabetes based on a case-control differential expression study performed using peripheral lymphomononuclear cells[11].

We discovered seven genes associated with CD (Figure 3): PIGC (1q24..33), FAAH (1p33), HLA-DQA1 (6pq21.32), HLA-B (6p21.33), PRKAB1 (12q24.23), LRRC37A2 (17q21.31), and HSBP1L1 (18q23). PIGC has been associated with inflammatory bowel disease as part of a larger Immunochip meta-analysis[24]. It is also significantly associated with T2D in our study. There has been no direct genetic evidence supporting the role of FAAH in CD, but recent experiments have shown that drugs targeting FAAH are effective against mouse models of colitis [40]. The HLA region has a known role in autoimmune disorders such as CD [13, 2]. PRKAB1 encodes the noncatalytic beta subunit of the AMP-activated protein kinase (AMPK), which has been previously experimentally validated to play an important role in IBD and is a therapeutic target for drugs used to treat CD and T2D [28]. Neither LRRC37A2 nor HSBP1L1 have been characterized in CD.

We found two genes associated with HT: SHMT1 and CIAO1. SHMT1 has previously been associated with hypertension, is a general prognostic marker for hypertension [34], is a marker for intra-cranial hypertension during space flight [45], and mediates the response to the angiogenesis inhibitor Bevacizumab[7]. CIAO1 encodes a protein in the iron-sulfur protein assembly complex and modulates activity of WT1, an oncogone associated with nephroblastomas [23]. It has not been studied in relation to HT.

We found multiple associations for RA in the MHC region: HLA-B, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB5, HLA-G, IER3, and MICA. Similarly, we found multiple MHC associations with T1D: BAK1, BTN3A2, HISTH1H2AG, HLA-B, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB5, HMGN4, and IER3. However, we discovered T1D associations with C1QTNF6, RBMS2, and RSP26. Larger meta-analysis of T1D has found association in the C1QTNF6 and RSP26 loci[12, 5]. However, RBSM2 has not been

previously linked to T1D.

Finally, we found four genes are associated with T2D: KHK, PIGC, POLR2J3, and TMEM158. Ketohexokinase (KHK) plays a role in fructose metabolism and has been studied extensively as a possible cause for T2D [10, 26]. A sub-threshold association in its locus has been identified in a recent study of T2D in a Japanese population [21]. The PIGC locus has been previously linked to BMI[39], but not to T2D. POLR2J3 also has not been studied in T2D previously.

We further characterized our TWAS associations by looking at overlaps with GWAS loci (Methods). For each significant gene, we looked for a GWAS hit within 1 MB of the gene body ($p < 5 \times 10^{-8}$), and found overlaps for 23 of 42 genes. We then asked how many of the 42 genes were later discovered by larger meta-analyses in the seven diseases using the NHGRI GWAS Catalog [18, 50] and found 10 TWAS genes overlap loci reported in the catalog. However, most of the common genes found by both TWAS and GWAS lie in the MHC region. We finally sought to validate our TWAS associations using orthogonal case-control expression data sets[22, 6, 52, 8, 48, 25].

Surprisingly, we found only 3 TWAS genes are differentially expressed between cases and controls (limma modified t-test[46], Benjamini-Hochberg FDR < 0.1). We investigated the ranking of the top 250 genes by TWAS and differential expression and found no significant overlap between the ranked lists (Methods). There are several possible explanations for this discrepancy, including the tissue in which expression was measured, sample size, and technical confounders.

4   Discussion

Transcriptome-wide association studies (TWAS) have proven to be a powerful approach for identifying new genes associated with a phenotype by cleverly combining reference expression data and available GWAS data. TWAS directly associate genes to disease, revealing new biological insights. Indeed, these ideas have been extended to additional levels of mediation, incorporating histone modification data to study genetic effects on epigenomic control of transcription [17]. However, the field has not yet fully appreciated the importance of uncertainty in these multi-stage regression models. We showed that the state-of-the-art methods for TWAS adequately control type I error, but lose power due to uncertainty. We proposed a novel two-stage Bayesian method, BAY-TS, which outperforms not only existing methods but also standard methods from both missing data and measurement error theory. In applications to seven diseases from WTCCC, our method identified new genes not identified by previous methods, which do not incorporate uncertainty.

Our results reveal that uncertainty arises from many sources, not only differences in the trained models of regulation. Using the same GWAS genotypes, different imputation pipelines did not yield the same gene associations. We showed that expression normalization has an impact on trained models used for TWAS, and demonstrated that models trained on regularized-log transformed data were better than those trained on published GTEx expression data. Other studies have shown inconsistency between different population of reference transcriptomes from the same tissue [16]. These differences can be attributed to technical variation, environmental noise, sequencing technology, processing pipelines, and population differences. There is a pressing need to develop methods which adequately account for all these sources of uncertainty.

6

5   Methods

5.1   Uncertainty in TWAS

We assume a continuous phenotype y_i with zero mean collected on n individuals, and regress phenotype on predicted expression w_i for each gene whose expression levels can be predicted from genotype information. To handle binary phenotypes, we estimate latent liabilities using LEAP[49] and regress predicted expression against individual liabilities. For ease of exposition, we describe a model with no additional covariates; these can be included as additional terms in the model with no modification to the algorithms.

$$y_i = \alpha w_i + \epsilon_i$$
$$y_i = \text{phenotype of individual } i$$
$$w_i = \text{predicted expression of individual } i$$
$$\epsilon_i = \text{error in equation}$$

Current models use only one prediction for w_i, based on a model of transcriptional regulation learned for the gene.

$$\hat{w}_i = G_i \hat{\beta}$$
$$G_i = \text{cis-regulatory SNP dosages}$$
$$\hat{\beta} = \text{eQTL effect size}$$

Here, we investigate models which account for the distribution of w_i and β. Assuming access to only one training cohort, we estimate these distributions by fitting k bootstrapped models regressing observed gene expression E on genotype G. Here, the regressions are performed using elastic net with regularization penalty tuned using cross-validation.

$$\hat{\beta}^{(k)} = \underset{\beta^{(k)}}{\arg\min} ||E - G\beta^{(k)}||^2 + \lambda_1 ||\beta^{(k)}||_1 + \lambda_2 ||\beta^{(k)}||_2$$

We estimate the first two moments of the distributions of w_i and β using the following equations:

7

$$\hat{w}_i^{(k)} = G\hat{\beta}^{(k)}$$

$$\bar{w}_i = \frac{1}{K}\sum_k w_i^{(k)}$$

$$\hat{\sigma}_{ui}^2 = \frac{1}{K-1}\sum_j (w_i^{(k)} - \bar{w}_i)^2$$

5.2 Bayesian two-stage regression model

To fit the Bayesian two-stage regression model for a gene (BAY-TS, Figure S1), we use the distributions of β_j learned using the k bootstrapped models for that gene as the prior in the second-stage regression.

$$P(\beta_j) = \mathcal{N}(\mathbb{E}[\beta_j], \mathbb{V}[\beta_j])$$
$$P(\alpha) = \mathcal{N}(0,1)$$
$$P(\epsilon_i) = \mathcal{N}(0, \sigma_e^2)$$
$$P(\sigma_e^2) = C^+(0, 10)$$

Here, $\mathcal{N}(\cdot, \cdot)$ denotes the Gaussian density and $C^+(\cdot, \cdot)$ denotes the half-Cauchy density. Our inference goal is to estimate a Bayes factor comparing the model described above to a null model where $\alpha = 0$. Rather than estimating the intractable model evidences and taking a ratio, we added a model indicator variable z and combined the null model $p(x,y) = N(0, \sigma_e^2)$ and the alternate model (described above). Then, the Bayes factor is given by the ratio $p(z=1)/p(z=0)$. We implemented the model using PyMC3 [38] and used the Metropolis-Hastings algorithm to perform inference. We ran the MCMC chain for 100,000 steps and used the last 10,000 samples to compute the Bayes factor.

5.3 Multiple imputation

To perform multiple imputation (MI), we fit the k bootstrapped models for w against the phenotype using linear regression, calculate an aggregated test statistic θ_{mi}, and compute a Wald test statistic.

$$\hat{\alpha}_{MI} = \frac{1}{K}\sum_k \alpha_k$$

$$\mathrm{Var}(\hat{\alpha}_{MI}) = \mathrm{Var}(\alpha) + \left(1 + \frac{1}{K}\right)\frac{1}{K}\sum_k \mathrm{Var}(\alpha_k)$$

$$\theta_{MI} = \frac{\hat{\alpha}_{MI}}{\mathrm{Var}(\hat{\alpha}_{MI})}$$

$$H_0 : \theta_{MI} = 0$$

$$H_1 : \theta_{MI} > 0$$

5.4   Regression calibration

We assume additive measurement error on the predicted expression value:

$$w_i = x_i + u_i$$

$$x_i = \text{true (latent) expression of individual } i$$

$$u_i = \text{error in predicted expression of individual } i$$

We assume measurement errors have zero mean and finite variance:

$$E[u_i] = 0$$

$$V[u_i] = \sigma_{ui}^2$$

To perform regression calibration (RC), we impute the true expression value and regress phenotype against this estimated true expression. Given $\hat{\sigma}_{ui}^2$, we regress y_i on $\hat{x}_i = \bar{w} - \hat{\kappa}(w_i - \bar{w})$, yielding estimate $\hat{\alpha}^*$. To estimate the association p-value, we perform a Wald test. We estimate Σ_α, the covariance of $\hat{\alpha}^*$, using a robust estimator [14]:

9

$$\Sigma_\alpha = M_{XX}^{-1} H M_{XX}^{-1}$$

$$M_{XX} = \frac{W'W}{n} - \hat{\Sigma}_u$$

$$W = n \times p \text{ design matrix (including intercept)}$$

$$\hat{\Sigma}_u = \text{covariance of measurement errors}$$

$$H = \frac{1}{n(n-p)} \sum_i \Delta_i \Delta_i'$$

$$r_i = y_i - W_i \hat{\alpha}$$

$$\Delta_i = W_i' r_i + \Sigma_{ui} \hat{\alpha}$$

$$\theta_{RC} = \left(\frac{\hat{\alpha}^*}{SE(\hat{\alpha}^*)} \right)^2 \sim \chi^2(\theta; 1)$$

$$H_0 : \theta_{RC} = 0$$

$$H_1 : \theta_{RC} > 0$$

5.5   Simulation study

We used imputed dosages for 4,884 samples from the Hypertension, 58C, and NBS cohorts as described below. We selected 193 genes with cis-heritable gene expression (likelihood ratio test, GREML) in all of three studies as previously reported[16]: Metabolic Syndrome in Men (METSIM), Netherlands Twin Registry (NTR), and Young Finns Study (YFS). We held out 350 individuals as the training cohort and used the rest as the test cohort.

For each gene, we sampled the causal fraction of eQTLs from (Single, 1%, 5%, 10%) of SNPs from the cis-regulatory window. We computed the genetic value of each individual $X = G\beta$ and added i.i.d. Gaussian noise to achieve proportion of variance explained (PVE) equal to 0.17 in expectation by sampling from $\mathcal{N}(0, \mathbb{V}[G\beta] * (1/.17 - 1))$, where $\mathbb{V}[G\beta]$ is the sample variance of the genetic values.

For each simulation, we sampled 40 causal genes and added i.i.d Gaussian noise to achieve PVE = 0.2 using the procedure described above. We computed the genetic value of each individual as $y = X\alpha$ and add Gaussian noise as described above. We evaluated the method using sample sizes of 5000 individuals. We also tested the performance varying the number of non-causal genes from 400 to 4000.

5.6   GWAS processing

We downloaded Affymetrix genotypes for the Wellcome Trust Case Control Consortium seven diseases study in OXSTATS format called using the Chiamo algorithm from the European Genome Archive. We downloaded probe identifiers, hg19 positions, and strand information (http://www.

`well.ox.ac.uk/~wrayner/strand/`) to convert positions to hg19 and used GTOOL version 0.7.5 to align all genotypes. We used PLINK version 1.09b to produce hard genotype calls with genotype probability threshold 0.99 and remove all SNPs and samples excluded from the original study.

We used SHAPEIT2 v2.r644 (ref. [19]) to exclude unalignable SNPs and phase the case and control cohorts independently for each autosome. We used default values for all model parameters. We used IMPUTE2 version 2.3.0 (ref. [20]) to impute into all SNPs and indels with MAF in European samples > 0.01. We divided the autosomes into 5 MB windows and threw out windows with fewer than 100 array probes.

For each cohort, we hard-called imputed dosages with genotype probability threshold 0.9. For each disease cohort, we produced a case-control set of hard-called genotypes for both the array genotypes and imputed genotypes by merging all chromosomes with the shared controls (1958 Birth Cohort and National Blood Services). We used GCTA 1.24 (ref. [53]) to estimate a genetic relatedness matrix on the case-control array genotypes and prune pairs of individuals with relatedness > 0.05. We used plink to remove these individuals from the imputed genotypes, and further remove indels and SNPs with missingness > 0.01, differential missingness ($p < 0.05$) or HWE $p < 10^{-5}$.

We used LEAP version 0.1.8.9 (ref. [49]) to estimate latent liabilities for each chromosome of each case-control dataset separately (holding out that chromosome) using the array genotypes. We used FastLMM version 0.2.26 (ref. [29]) to compute association p-values for the imputed genotypes using kinship matrices estimated from the array genotypes (described above). We made extensive use of GNU parallel[47] to facilitate the analysis.

5.7   Reference expression processing

We downloaded genotypes and RNA-Seq read counts in the v6 release of GTEx from dbGaP. We restricted our analysis to only those genes which had RNASeqC gene-level read count $>= 10$ in at least 10 individuals, resulting in 12,049 genes. We transformed the counts to regularized log-transform values (RLog) using DESEq2, adjusting for sequencing depth using the median-of-ratios method [32]. We trained our models on these normalized values to models trained on the published expression values (GTEX-Norm).

We extracted genotypes for SNPs within 500kb upstream and downstream of the transcription start and end sites for each gene using plink. We filtered sites with missingness > 0.01 or HWE $p < 10^{-5}$. For models fit on the published expression values (GTEx-Norm), we included 3 genotype principal components (PCs), 35 expression PEER factors, gender, and sequencing platform as covariates. For models fit on RLog, we used the same covariates but used 10 expression PCs instead of the PEER factors.

To find the optimal elastic net penalty parameter and l1/l2 regularization ratio, we used Elastic-NetCV from the scikit-learn package [36] with possible l1 ratio values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.95, and 0.99 and 100 uniformly distributed penalty parameters from 0.1 to 1. Using the best fitted penalty parameters parameters, we fit 50 bootstrapped models by sampling 300 individuals from the 338 samples.

We imputed expression using PrediXcan software as well as an independent implementation. In our independent implementation, we filtered all GWAS sites that had higher than 10% missing

genotypes. Additionally, we assigned the average value to all missing genotypes (which is possible for best-guess imputed genotypes after thresholding on the posterior probability of any genotype call). We note that the implementation of PrediXcan assumes there is no missingness in the data.

To calculate cross-validation prediction accuracy, we predicted hold-out gene expression using only genotype (omitting covariates). We used 5-fold cross-validation and calculated the average R^2 across the folds.

We define the variance ratio (VR) as the ratio of variance of mean imputed expression to the mean of the variance of imputed expression. Intuitively, the VR compares the within-individual variation in imputed expression to between-individual variation. We estimate VR by estimating the necessary means and variances over the 50 bootstrapped models described above.

5.8   GEO data

We downloaded case-control expression datasets for the seven diseases from GEO and processed them using the GEO2R web service: GSE12654 (BD) [22], GSE20681 (CAD) [6], GSE6731 (CD) [52], GSE703 (HT) [8], GSE15573 (RA) [48], GSE55100 (T1D) [54], GSE21321 (T2D) [25]. For each data set, we found differentially expressed genes between cases and controls using limma and the GEO2R web service. We assessed significance in the ranking between TWAS lists and GEO lists using the R package OrderedList [55], restricting to the top 250 overlapping genes for each disease.
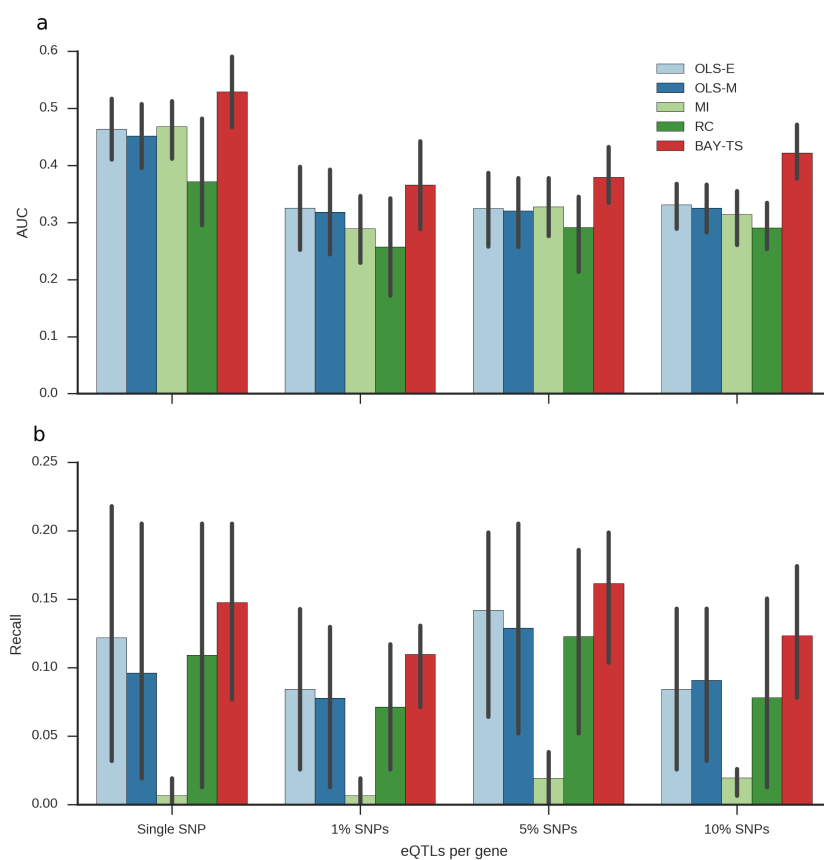
12

Figure 1: Simulation results. a) Area under precision-recall curve (AUPRC) for each method. Error bars represent standard error of AUPRC estimated over 4 replicates. b) Recall of each method controlling FDR at 10%. Error bars represent standard error of recall over 4 replicates.
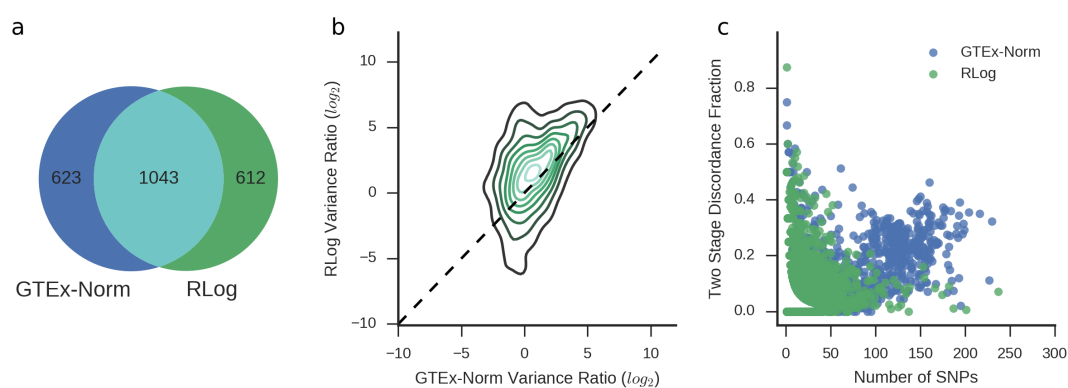
Figure 2: Impact of expression normalization on trained models of transcriptional regulation. a) The number of genes with cross-validation $R^2 \geq 0$. b) Density of the distribution of variance ratio (VR) for each of the 1043 common genes with cross-validation $R^2 \geq 0$. Contours denote surfaces of equal density. c) The total number of selected predictors per gene and the fraction of discordant predictors selected per gene.

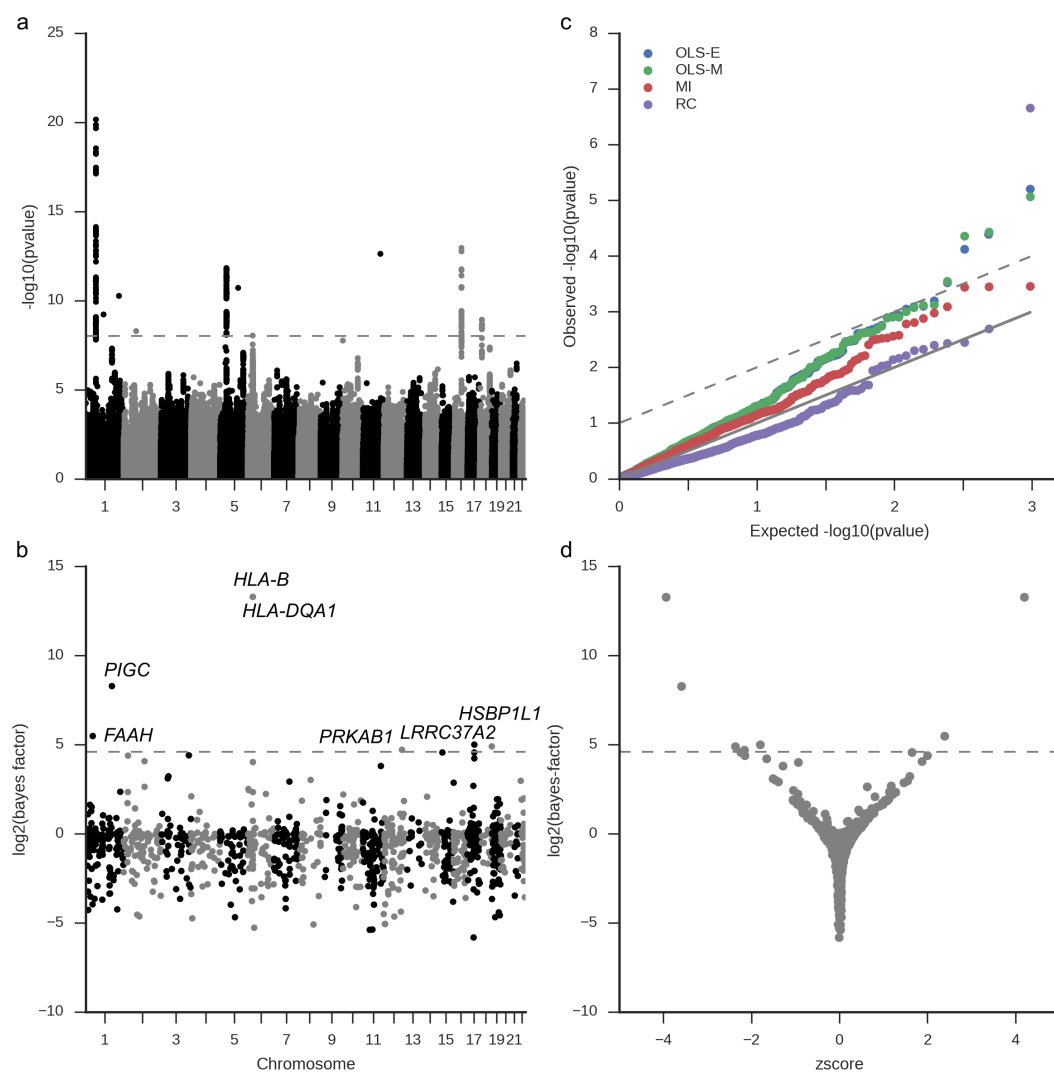Figure 3: TWAS results for Crohn's Disease. a) Manhattan plot of GWAS summary statistics estimated using FastLMM and LEAP on the WTCCC GWAS cohort. b) BAY-TS TWAS on the same samples reveals 7 genes that are significantly associated with Crohn's Disease. c) Quantile-quantile plot of TWAS test statistics for frequentist methods. d) Bayes factor and z-score of BAY-TS associations. A Bayes factor of at least 24 corresponds to FDR 10%.

Table 1: TWAS results for Bayesian two-stage regression model. BF: Bayes factor of BAY-TS. We only report genes with BF > 24, corresponding for FDR 10%. Effect size: Estimate of gene effect on disease liability. R^2: cross-validation prediction accuracy. Ratio: variance ratio, the ratio of within-individual variance in imputed expression to across-individual variance. GWAS: best p-value in the WTCCC cohort within 1MB of the gene body. Catalog: best p-value reported in the NHGRI GWAS catalog within 1MB of the gene body. GEO: p-value for differential expression in independent case-control expression cohort

Gene	Disease	Chr	BF	Effect size	R2	Ratio	GWAS	Catalog	Diff. expr
SARDH	BD	9	76.52	-0.43	0.03	156.52	7.71e-04	-	0.98
ZNF79	BD	9	25.18	-1.01	0.04	23.03	1.40e-03	-	0.38
WDR25	BD	14	28.07	-0.69	0.15	41.10	6.76e-04	-	-
PARL	CAD	3	46.39	0.38	0.03	136.61	1.02e-04	-	0.26
TMEM158	CAD	3	48.50	-1.03	0.01	1.82	2.09e-03	-	0.092
CYTH3	CAD	7	28.15	-1.47	0.02	0.71	1.42e-04	-	0.47
DAGLB	CAD	7	77.12	2.13	0.09	4.13	1.42e-04	-	0.39
FAAH	CD	1	44.66	0.62	0.01	8.39	1.86e-04	-	0.38
PIGC	CD	1	311.50	-1.73	0.12	34.47	5.93e-08	3e-22	0.0075
HLA-B	CD	6	9999.00	-0.76	0.02	33.40	9.22e-09	7e-32	0.34
HLA-DQA1	CD	6	10000.00	0.59	0.39	27.64	6.12e-08	9e-59	0.98
PRKAB1	CD	12	26.17	-0.87	0.03	74.76	1.33e-03	-	0.34
LRRC37A2	CD	17	32.00	-0.47	0.34	47.37	3.55e-05	-	0.086
HSBP1L1	CD	18	29.86	-0.88	0.25	32.46	3.05e-05	-	-
CIAO1	HT	2	49.00	1.07	0.03	48.08	6.04e-05	-	-
SHMT1	HT	17	9999.00	0.47	0.15	45.07	1.48e-04	-	-
HLA-B	RA	6	311.50	-1.65	0.02	16.04	9.54e-33	3e-10	0.45
HLA-DQA1	RA	6	10000.00	3.90	0.39	8.21	2.25e-86	1e-299	-
HLA-DQA2	RA	6	10000.00	1.00	0.37	16.46	2.25e-86	1e-299	0.29
HLA-DQB1	RA	6	10000.00	4.97	0.61	6.58	2.25e-86	1e-299	0.85
HLA-DQB2	RA	6	10000.00	3.16	0.37	12.05	2.25e-86	1e-299	-
HLA-DRB1	RA	6	10000.00	-1.20	0.26	15.60	2.25e-86	1e-299	0.33
HLA-DRB5	RA	6	10000.00	-3.14	0.62	8.28	2.25e-86	1e-299	0.26
HLA-G	RA	6	4999.00	-0.30	0.34	24.16	4.51e-16	-	0.75
IER3	RA	6	10000.00	1.55	0.05	16.03	9.13e-12	-	0.0089
MICA	RA	6	10000.00	0.94	0.29	62.73	9.54e-33	3e-10	0.27
BAK1	T1D	6	10000.00	-1.10	0.14	32.53	2.35e-31	-	0.21
BTN3A2	T1D	6	10000.00	-0.47	0.19	148.71	1.94e-13	-	0.44
HIST1H2AG	T1D	6	10000.00	-3.33	0.02	3.26	2.26e-15	-	-
HLA-B	T1D	6	10000.00	-4.84	0.02	2.24	1.44e-91	-	0.37
HLA-DQA1	T1D	6	10000.00	5.82	0.39	2.23	0.00e+00	5e-134	0.045
HLA-DQA2	T1D	6	10000.00	9.79	0.37	5.93	0.00e+00	5e-134	-
HLA-DQB1	T1D	6	10000.00	8.49	0.61	1.34	0.00e+00	5e-134	0.00018
HLA-DQB2	T1D	6	10000.00	7.40	0.37	26.58	0.00e+00	5e-134	0.11
HLA-DRB1	T1D	6	10000.00	-3.97	0.26	8.23	0.00e+00	5e-134	0.24
HLA-DRB5	T1D	6	554.56	-2.08	0.62	0.46	0.00e+00	5e-134	-
HMGN4	T1D	6	10000.00	-1.44	0.09	9.59	1.94e-13	-	0.22
IER3	T1D	6	10000.00	5.08	0.05	15.09	1.16e-48	-	0.13
RBMS2	T1D	12	10000.00	-3.31	0.23	1.23	8.93e-12	-	0.026
RPS26	T1D	12	10000.00	-3.10	0.02	12.37	8.93e-12	2e-25	-
C1QTNF6	T1D	22	27.82	0.83	0.06	45.02	1.17e-05	2e-08	0.73
PIGC	T2D	1	33.01	-0.98	0.12	30.33	1.73e-05	-	-
KHK	T2D	2	28.15	-0.60	0.04	36.63	8.13e-04	2e-06	0.087
TMEM158	T2D	3	103.17	-1.06	0.01	1.77	9.53e-04	-	0.2
POLR2J3	T2D	7	624.00	-0.40	0.08	4.27	1.13e-03	-	-

16

References

[1] Elif F. Acar and Lei Sun. A Generalized Kruskal-Wallis Test Incorporating Group Uncertainty with Application to Genetic Association Studies. *Biometrics*, 69(2):427–435, jun 2013.

[2] Tariq Ahmad, Alessandro Armuzzi, Mike Bunce, Kim Mulcahy–Hawes, Sara E. Marshall, Timothy R. Orchard, Jonathan Crawshaw, Oliver Large, Arjuna De Silva, Julia T. Cook, Martin Barnardo, Sue Cullen, Ken I. Welsh, and Derek P. Jewell. The molecular classification of the clinical manifestations of Crohn's disease. *Gastroenterology*, 122(4):854–866, 2002.

[3] Mariet Allen, Minerva M. Carrasquillo, Cory Funk, Benjamin D. Heavner, Fanggeng Zou, Curtis S. Younkin, Jeremy D. Burgess, High-Seng Chai, Julia Crook, James A. Eddy, Hongdong Li, Ben Logsdon, Mette A. Peters, Kristen K. Dang, Xue Wang, Daniel Serie, Chen Wang, Thuy Nguyen, Sarah Lincoln, Kimberly Malphrus, Gina Bisceglio, Ma Li, Todd E. Golde, Lara M. Mangravite, Yan Asmann, Nathan D. Price, Ronald C. Petersen, Neill R. Graff-Radford, Dennis W. Dickson, Steven G. Younkin, Nilüfer Ertekin-Taner, M. M. Carrasquillo, D. Harold, J. C. Lambert, S. Seshadri, A. C. Naj, P. Hollingworth, J. C. Lambert, G. U. Hoglinger, J. Simon-Sanchez, F. Zou, A. L. Dixon, V. Emilsson, A. J. Saykin, F. Zou, M. Allen, M. Allen, M. Allen, M Allen, A. J. Myers, J. A. Webster, J. Chapuis, L. N. Hazrati, A. Ramasamy, S. B. Montgomery, J. M. Derry, M. Allen, G McKhann, LA Farrer, H. Braak, E. Braak, J. J. Hauw, S. S. Mirra, J. Wang, D. W. Dickson, J. Q. Trojanowski, V. M. Lee, P. Du, W. A. Kibbe, S. M. Lin, S. Purcell, A. T. Magis, C. C. Funk, N. D. Price, and C Funk. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Scientific Data*, 3:160089, oct 2016.

[4] K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, T. Esko, W. Winckler, J. N. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shabalin, G. Li, Y.-H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. M. Goldmann, D. Koller, R. Guigo, M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, M. T. Moser, B. M. Gillard, E. Karasik, K. Ramsey, C. Choi, B. A. Foster, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. M. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. D. Jewell, P. A. Branton, L. H. Sobin, M. Barcus, L. Qi, J. McLean, P. Hariharan, K. S. Um, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, and E. T. Dermitzakis. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, may 2015.

[5] Jeffrey C Barrett, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Cécile Julier, Grant Morahan, Jørn Nerup, Concepcion Nierras, Vincent Plagnol, Flemming Pociot, Helen Schuilenburg, Deborah J Smyth, Helen Stevens, John A Todd, Neil M Walker, Stephen S Rich, and Type 1 Diabetes Genetics Consortium. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, 41(6):703–707, jun 2009.

[6] Philip Beineke, Karen Fitch, Heng Tao, Michael R Elashoff, Steven Rosenberg, William E Kraus, James A Wingrove, and PREDICT Investigators. A whole blood gene expression-based signature for smoking status. *BMC medical genomics*, 5:58, dec 2012.

[7] Barna Budai, Viktor Komlósi, Vilmos Adleff, Éva Pap, Andrea Réti, Tünde Nagy, Judit Kralovánszky, István Láng, and Erika Hitre. Impact of SHMT1 polymorphism on the clinical outcome of patients with metastatic colorectal cancer treated with first-line FOLFIRI+bevacizumab. *Pharmacogenetics and Genomics*, 22(1):69–72, jan 2012.

[8] Todd M Bull, Christopher D Coldren, Mark Moore, Sylk M Sotto-Santiago, David V Pham, S Patrick Nana-Sinkam, Norbert F Voelkel, and Mark W Geraci. Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension. *American journal of respiratory and critical care medicine*, 170(8):911–9, oct 2004.

[9] Paul R. Burton, David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, Mark I. McCarthy, Willem H. Ouwehand, Nilesh J. Samani, John A. Todd, Peter Donnelly, Jeffrey C. Barrett, Paul R. Burton, Dan Davison, Peter Donnelly, Doug Easton, David Evans, Hin-Tak Leung, Jonathan L. Marchini, Andrew P. Morris, Chris C. A. Spencer, Martin D. Tobin, Lon R. Cardon, David G. Clayton, Antony P. Attwood, James P. Boorman, Barbara Cant, Ursula Everson, Judith M. Hussey, Jennifer D. Jolley, Alexandra S. Knight, Kerstin Koch, Elizabeth Meech, Sarah Nutland, Christopher V. Prowse, Helen E. Stevens, Niall C. Taylor, Graham R. Walters, Neil M. Walker, Nicholas A. Watkins, Thilo Winzer, John A. Todd, Willem H. Ouwehand, Richard W. Jones, Wendy L. McArdle, Susan M. Ring, David P. Strachan, Marcus Pembrey, Gerome Breen, David St Clair, Sian Caesar, Katherine Gordon-Smith, Lisa Jones, Christine Fraser, Elaine K. Green, Detelina Grozeva, Marian L. Hamshere, Peter A. Holmans, Ian R. Jones, George Kirov, Valentina Moskvina, Ivan Nikolov, Michael C. O'Donovan, Michael J. Owen, Nick Craddock, David A. Collier, Amanda Elkin, Anne Farmer, Richard Williamson, Peter McGuffin, Allan H. Young, I. Nicol Ferrier, Stephen G. Ball, Anthony J. Balmforth, Jennifer H. Barrett, D. Timothy Bishop, Mark M. Iles, Azhar Maqbool, Nadira Yuldasheva, Alistair S. Hall, Peter S. Braund, Paul R. Burton, Richard J. Dixon, Massimo Mangino, Suzanne Stevens, Martin D. Tobin, John R. Thompson, Nilesh J. Samani, Francesca Bredin, Mark Tremelling, Miles Parkes, Hazel Drummond, Charles W. Lees, Elaine R. Nimmo, Jack Satsangi, Sheila A. Fisher, Alastair Forbes, Cathryn M. Lewis, Clive M. Onnie, Natalie J. Prescott, Jeremy Sanderson, Christopher G. Mathew, Jamie Barbour, M. Khalid Mohiuddin, Catherine E. Todhunter, John C. Mansfield, Tariq Ahmad, Fraser R. Cummings, Derek P. Jewell, John Webster, Morris J. Brown, David G. Clayton, G. Mark Lathrop, John Connell, Anna Dominiczak, Nilesh J. Samani, Carolina A. Braga Marcano, Beverley Burke, Richard Dobson, Johannie Gungadoo, Kate L. Lee, Patricia B. Munroe, Stephen J. Newhouse, Abiodun Onipinla, Chris Wallace, Mingzhan Xue, Mark Caulfield, Martin Farrall, Anne Barton, The Biologics in RA Genetics and Genomics (BRAGGS), Ian N. Bruce, Hannah Donovan, Steve Eyre, Paul D. Gilbert,

18

Samantha L. Hider, Anne M. Hinks, Sally L. John, Catherine Potter, Alan J. Silman, Deborah P. M. Symmons, Wendy Thomson, Jane Worthington, David G. Clayton, David B. Dunger, Sarah Nutland, Helen E. Stevens, Neil M. Walker, Barry Widmer, John A. Todd, Timothy M. Frayling, Rachel M. Freathy, Hana Lango, John R. B. Perry, Beverley M. Shields, Michael N. Weedon, Andrew T. Hattersley, Graham A. Hitman, Mark Walker, Kate S. Elliott, Christopher J. Groves, Cecilia M. Lindgren, Nigel W. Rayner, Nicholas J. Timpson, Eleftheria Zeggini, Mark I. McCarthy, Melanie Newport, Giorgio Sirugo, Emily Lyons, Fredrik Vannberg, Adrian V. S. Hill, Linda A. Bradbury, Claire Farrar, Jennifer J. Pointon, Paul Wordsworth, Matthew A. Brown, Jayne A. Franklyn, Joanne M. Heward, Matthew J. Simmonds, Stephen C. L. Gough, Sheila Seal, Breast Cancer Susceptibility Collaboration (UK), Michael R. Stratton, Nazneen Rahman, Maria Ban, An Goris, Stephen J. Sawcer, Alastair Compston, David Conway, Muminatou Jallow, Melanie Newport, Giorgio Sirugo, Kirk A. Rockett, Dominic P. Kwiatkowski, Suzannah J. Bumpstead, Amy Chaney, Kate Downes, Mohammed J. R. Ghori, Rhian Gwilliam, Sarah E. Hunt, Michael Inouye, Andrew Keniry, Emma King, Ralph McGinnis, Simon Potter, Rathi Ravindrarajah, Pamela Whittaker, Claire Widden, David Withers, Panos Deloukas, Hin-Tak Leung, Sarah Nutland, Helen E. Stevens, Neil M. Walker, John A. Todd, Doug Easton, David G. Clayton, Paul R. Burton, Martin D. Tobin, Jeffrey C. Barrett, David Evans, Andrew P. Morris, Lon R. Cardon, Niall J. Cardin, Dan Davison, Teresa Ferreira, Joanne Pereira-Gale, Ingileif B. Hallgrimsdóttir, Bryan N. Howie, Jonathan L. Marchini, Chris C. A. Spencer, Zhan Su, Yik Ying Teo, Damjan Vukcevic, Peter Donnelly, David Bentley, Matthew A. Brown, Lon R. Cardon, Mark Caulfield, David G. Clayton, Alistair Compston, Nick Craddock, Panos Deloukas, Peter Donnelly, Martin Farrall, Stephen C. L. Gough, Alistair S. Hall, Andrew T. Hattersley, Adrian V. S. Hill, Dominic P. Kwiatkowski, Christopher G. Mathew, Mark I. McCarthy, Willem H. Ouwehand, Miles Parkes, Marcus Pembrey, Nazneen Rahman, Nilesh J. Samani, Michael R. Stratton, John A. Todd, and Jane Worthington. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, jun 2007.

[10] Pietro Cirillo, Michael S Gersch, Wei Mu, Philip M Scherer, Kyung Mee Kim, Loreto Gesualdo, George N Henderson, Richard J Johnson, and Yuri Y Sautin. Ketohexokinase-dependent metabolism of fructose induces proinflammatory mediators in proximal tubular cells. *Journal of the American Society of Nephrology : JASN*, 20(3):545–53, mar 2009.

[11] C. V. A. Collares, A. F. Evangelista, D. J. Xavier, P. Takahashi, R. Almeida, C. Macedo, F. Manoel-Caetano, M. C. Foss, M. C. Foss-Freitas, D. M. Rassi, E. T. Sakamoto-Hojo, G. A. Passos, and E. A. Donadi. Transcriptome meta-analysis of peripheral lymphomononuclear cells indicates that gestational diabetes is closer to type 1 diabetes than to type 2 diabetes mellitus. *Molecular Biology Reports*, 40(9):5351–5358, sep 2013.

[12] Jason D Cooper, Deborah J Smyth, Adam M Smiles, Vincent Plagnol, Neil M Walker, James E Allen, Kate Downes, Jeffrey C Barrett, Barry C Healy, Josyf C Mychaleckyj, James H Warram, and John A Todd. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics*, 40(12):1399–1401, dec 2008.

[13] Sheila A Fisher, Mark Tremelling, Carl A Anderson, Rhian Gwilliam, Suzannah Bumpstead, Natalie J Prescott, Elaine R Nimmo, Dunecan Massey, Carlo Berzuini, Christopher Johnson, Jeffrey C Barrett, Fraser R Cummings, Hazel Drummond, Charlie W Lees, Clive M Onnie, Catherine E Hanson, Katarzyna Blaszczyk, Mike Inouye, Philip Ewels, Radhi Ravindrarajah,

19

Andrew Keniry, Sarah Hunt, Martyn Carter, Nick Watkins, Willem Ouwehand, Cathryn M Lewis, Lon Cardon, Alan Lobo, Alastair Forbes, Jeremy Sanderson, Derek P Jewell, John C Mansfield, Panos Deloukas, Christopher G Mathew, Miles Parkes, and Jack Satsangi. Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. *Nature Genetics*, 40(6):710–712, jun 2008.

[14] Wayne A. Fuller. *Measurement Error Models*. John Wiley & Sons, Inc., 1987.

[15] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, aug 2015.

[16] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, feb 2016.

[17] Alexander Gusev, Nick Mancuso, Hilary K Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Edwin Oh, Steven McCaroll, Benjamin Neale, Roel Ophoff, Michael C O'Donovan, Nicholas Katsanis, Gregory E Crawford, Patrick F Sullivan, Bogdan Pasaniuc, and Alkes L Price. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *bioRxiv*, 2016.

[18] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, jun 2009.

[19] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8):955–959, jul 2012.

[20] Bryan N. Howie, Peter Donnelly, Jonathan Marchini, JD Rioux, RJ Xavier, KD Taylor, MS Silverberg, P Goyette, J Gudmundsson, P Sulem, A Manolescu, LT Amundadottir, D Gudbjartsson, KA Frazer, DG Ballinger, DR Cox, DA Hinds, LL Stuve, B Servin, M Stephens, J Marchini, B Howie, S Myers, G McVean, P Donnelly, DY Lin, Y Hu, BE Huang, DL Nicolae, E Zeggini, LJ Scott, R Saxena, BF Voight, JL Marchini, JC Barrett, S Hansoul, DL Nicolae, JH Cho, RH Duerr, P Scheet, M Stephens, SR Browning, BL Browning, BL Browning, SR Browning, JC Barrett, DG Clayton, P Concannon, B Akolkar, JD Cooper, S Purcell, B Neale, K Todd-Brown, L Thomas, MA Ferreira, N Li, M Stephens, S Myers, L Bottolo, C Freeman, G McVean, P Donnelly, J Wakeley, A Kong, G Masson, ML Frigge, A Gylfason, P Zusmanovich, Y Guan, M Stephens, J Marchini, D Cutler, N Patterson, M Stephens, E Eskin, W Chen, Y Li, G Abecasis, Z Zhao, N Timofeev, SW Hartley, DH Chui, S Fucharoen, L Huang, Y Li, AB Singleton, JA Hardy, and G Abecasis. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, 5(6):e1000529, jun 2009.

[21] Minako Imamura, Atsushi Takahashi, Toshimasa Yamauchi, Kazuo Hara, Kazuki Yasuda, Niels Grarup, Wei Zhao, Xu Wang, Alicia Huerta-Chagoya, Cheng Hu, Sanghoon Moon, Jirong Long, Soo Heon Kwak, Asif Rasheed, Richa Saxena, Ronald C. W. Ma, Yukinori Okada, Minoru Iwata, Jun Hosoe, Nobuhiro Shojima, Minaka Iwasaki, Hayato Fujita, Ken Suzuki, John Danesh, Torben Jørgensen, Marit E. Jørgensen, Daniel R. Witte, Ivan Brandslund, Cramer Christensen, Torben Hansen, Josep M. Mercader, Jason Flannick, Hortensia Moreno-Macías, Noël P. Burtt, Rong Zhang, Young Jin Kim, Wei Zheng, Jai Rup Singh, Claudia H. T. Tam, Hiroshi Hirose, Hiroshi Maegawa, Chikako Ito, Kohei Kaku, Hirotaka Watada, Yasushi Tanaka, Kazuyuki Tobe, Ryuzo Kawamori, Michiaki Kubo, Yoon Shin Cho, Juliana C. N. Chan, Dharambir Sanghera, Philippe Frossard, Kyong Soo Park, Xiao-Ou Shu, Bong-Jo Kim, Jose C. Florez, Teresa Tusié-Luna, Weiping Jia, E Shyong Tai, Oluf Pedersen, Danish Saleheen, Shiro Maeda, and Takashi Kadowaki. Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nature Communications*, 7:10531, jan 2016.

[22] K Iwamoto, C Kakiuchi, M Bundo, K Ikeda, and T Kato. Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Molecular psychiatry*, 9(4):406–16, apr 2004.

[23] RW Johnstone, J Wang, N Tommerup, and H Vissing. Ciao 1 is a novel WD40 protein that interacts with the tumor suppressor protein WT1. *Journal of Biological*, 1998.

[24] Luke Jostins, Stephan Ripke, Rinse K. Weersma, Richard H. Duerr, Dermot P. McGovern, Ken Y. Hui, James C. Lee, L. Philip Schumm, Yashoda Sharma, Carl A. Anderson, Jonah Essers, Mitja Mitrovic, Kaida Ning, Isabelle Cleynen, Emilie Theatre, Sarah L. Spain, Soumya Raychaudhuri, Philippe Goyette, Zhi Wei, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Leila Amininejad, Ashwin N. Ananthakrishnan, Vibeke Andersen, Jane M. Andrews, Leonard Baidoo, Tobias Balschun, Peter A. Bampton, Alain Bitton, Gabrielle Boucher, Stephan Brand, Carsten Büning, Ariella Cohain, Sven Cichon, Mauro D'Amato, Dirk De Jong, Kathy L. Devaney, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Lynnette R. Ferguson, Denis Franchimont, Karin Fransen, Richard Gearry, Michel Georges, Christian Gieger, Jürgen Glas, Talin Haritunians, Ailsa Hart, Chris Hawkey, Matija Hedl, Xinli Hu, Tom H. Karlsen, Limas Kupcinskas, Subra Kugathasan, Anna Latiano, Debby Laukens, Ian C. Lawrance, Charlie W. Lees, Edouard Louis, Gillian Mahy, John Mansfield, Angharad R. Morgan, Craig Mowat, William Newman, Orazio Palmieri, Cyriel Y. Ponsioen, Uros Potocnik, Natalie J. Prescott, Miguel Regueiro, Jerome I. Rotter, Richard K. Russell, Jeremy D. Sanderson, Miquel Sans, Jack Satsangi, Stefan Schreiber, Lisa A. Simms, Jurgita Sventoraityte, Stephan R. Targan, Kent D. Taylor, Mark Tremelling, Hein W. Verspaget, Martine De Vos, Cisca Wijmenga, David C. Wilson, Juliane Winkelmann, Ramnik J. Xavier, Sebastian Zeissig, Bin Zhang, Clarence K. Zhang, Hongyu Zhao, Mark S. Silverberg, Vito Annese, Hakon Hakonarson, Steven R. Brant, Graham Radford-Smith, Christopher G. Mathew, John D. Rioux, Eric E. Schadt, Mark J. Daly, Andre Franke, Miles Parkes, Severine Vermeire, Jeffrey C. Barrett, Judy H Cho, and Judy H Cho. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, oct 2012.

[25] Dwi Setyowati Karolina, Arunmozhiarasi Armugam, Subramaniam Tavintharan, Michael T K Wong, Su Chi Lim, Chee Fang Sum, and Kandiah Jeyaseelan. MicroRNA 144 impairs insulin

signaling by inhibiting the expression of insulin receptor substrate 1 in type 2 diabetes mellitus. *PloS one*, 6(8):e22839, 2011.

[26] Zeid Khitan and Dong Hyun Kim. Fructose: a key factor in the development of metabolic syndrome and hypertension. *Journal of nutrition and metabolism*, 2013:682673, 2013.

[27] H M Lachman. Copy variations in schizophrenia and bipolar disorder. *Cytogenetic and genome research*, 123(1-4):27–35, 2008.

[28] Seon-Yeong Lee, Seung Hoon Lee, Eun-Ji Yang, Eun-Kyung Kim, Jae-Kyung Kim, Dong-Yun Shin, Mi-La Cho, DC Baumgart, WJ Sandborn, A Dignass, R Eliakim, F Magro, C Maaser, Y Chowers, K Geboes, A Raza, W Yousaf, R Giannella, MT Shata, J Galvez, CJ Wruck, A Fragoulis, A Gurzynski, LO Brandenburg, YW Kan, K Chan, FM Gu, QL Li, Q Gao, JH Jiang, K Zhu, XY Huang, XO Yang, AD Panopoulos, R Nurieva, SH Chang, D Wang, SS Watowich, EJ Yang, J Lee, SY Lee, EK Kim, YM Moon, YO Jung, J Jhun, J Lee, JK Byun, EK Kim, JW Woo, JH Lee, H Capiralla, V Vingtdeux, J Venkatesh, U Dreses-Werringloer, H Zhao, P Davies, C Kitabayashi, T Fukada, M Kanamoto, W Ohashi, S Hojyo, T Atsumi, L Zhou, JE Lopes, MM Chong, II Ivanov, R Min, GD Victora, MS Maddur, P Miossec, SV Kaveri, J Bayry, C Selmi, CS Park, BR Bang, HS Kwon, KA Moon, TB Kim, KY Lee, HA Hirsch, D Iliopoulos, K Struhl, K Isoda, JL Young, A Zirlik, LA MacFarlane, N Tsuboi, N Gerdes, A Hahn-Windgassen, V Nogueira, CC Chen, JE Skeen, N Sonenberg, N Hay, DB Shackelford, RJ Shaw, HJ Son, J Lee, SY Lee, EK Kim, MJ Park, KW Kim, DA van Heel, IA Udalova, AP De Silva, DP McGovern, Y Kinouchi, J Hull, I Marafini, F Zorzi, S Codazza, F Pallone, G Monteleone, L Deng, JF Zhou, RS Sellers, JF Li, AV Nguyen, Y Wang, ME Himmel, Y Yao, PC Orban, TS Steiner, MK Levings, N Eastaff-Leung, N Mabarrack, A Barbour, A Cummins, S Barry, H Yu, D Pardoll, R Jove, R Izutani, EY Loh, HC Reinecker, Y Ohno, RD Fusunyan, GR Lichtenstein, HC Reinecker, M Steffen, T Witthoeft, I Pflueger, S Schreiber, RP MacDermott, F Scaldaferri, S Vetrano, M Sans, V Arena, G Straface, E Stigliano, J Zhou, J Wulfkuhle, H Zhang, P Gu, Y Yang, J Deng, D Micic, G Cvijovic, V Trajkovic, LH Duntas, S Polovina, Y Zheng, AY Rudensky, P Miossec, JK Kolls, WA Goodman, AB Young, TS McCormick, KD Cooper, A Levine, L Passerini, SE Allan, M Battaglia, S Di Nunzio, AN Alstad, MK Levings, SJ Koh, JM Kim, IK Kim, SH Ko, and JS Kim. Metformin Ameliorates Inflammatory Bowel Disease by Suppression of the STAT3 Signaling Pathway and Regulation of the between Th17/Treg Balance. *PLOS ONE*, 10(9):e0135858, sep 2015.

[29] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, sep 2011.

[30] Roderick J. A. Little and Rubin B. Donald. *Statistical Analysis with Missing Data*. Wiley Online Library, 2002.

[31] Jing Liu, Xiao-feng Huang, Ju-xiang Liu, Jin-xing Quan, Li-min Tian, Xiao-juan Huang, Jia Liu, Yan-jia Xu, Qi Zhang, Shu-lan Zhang, Xiao-hui Chen, and Rui-lan Niu. Association of PARL Gene Rs3732581, Rs73887537 Polymorphisms with Type 2 Diabetes Mellitus, Insulin Resistance and Blood Lipid Levels in Chinese Population. *Journal of Metabolic Syndrome*, 03(01):1–7, 2014.

[32] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, dec 2014.

[33] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913, jul 2007.

[34] Tobore Onojighofia, Natasha Anand, Brian Meshkin, Sanford Silverman, Derrick Holman, John Hubbard, May Hafez, and Svetlana Kantorovich. Abstract P156: A Novel Gene-based Tool to Predict the Risk of Essential Hypertension and Initial Validation. *Hypertension*, 66(Suppl 1), 2015.

[35] Cameron Palmer, Itsik Pe'er, SR Browning, J Marchini, B Howie, S Myers, G McVean, P Donnelly, AL Price, A Tandon, N Patterson, KC Barnes, N Rafaels, I Ruczinski, HM Kang, JH Sul, SK Service, NA Zaitlen, Sy Kong, NB Freimer, S Prabhu, I Pe'er, AL Price, NJ Patterson, RM Plenge, ME Weinblatt, NA Shadick, D Reich, GR Abecasis, SS Cherny, WO Cookson, LR Cardon, BM Neale, MA Rivas, BF Voight, D Altshuler, B Devlin, M Orho-Melander, RJA Little, DB Rubin, YF Pei, J Li, L Zhang, CJ Papasian, HW Deng, BL Browning, SR Browning, B Howie, C Fuchsberger, M Stephens, J Marchini, GR Abecasis, M Nothnagel, D Ellinghaus, S Schreiber, M Krawczak, A Franke, K Nho, L Shen, S Kim, S Swaminathan, SL Risacher, AJ Saykin, AE Locke, B Kahali, SI Berndt, AE Justice, TH Pers, FR Day, AR Wood, T Esko, J Yang, S Vedantam, TH Pers, S Gustafsson, EF Acar, L Sun, Y Li, CJ Willer, J Ding, P Scheet, GR Abecasis, BN Howie, P Donnelly, J Marchini, C Fuchsberger, GR Abecasis, DA Hinds, CJ Willer, Y Li, GR Abecasis, O Harel, JL Schafer, JW Graham, AE Olchowski, TD Gilreath, D Lee, TB Bigdeli, BP Riley, AH Fanous, SA Bacanu, B Pasaniuc, N Zaitlen, H Shi, G Bhatia, A Gusev, J Pickrell, D Lee, TB Bigdeli, VS Williamson, VI Vladimirov, BP Riley, AH Fanous, MD Mailman, M Feolo, Y Jin, M Kimura, K Tryka, R Bagoutdinov, S Purcell, B Neale, K Todd-Brown, L Thomas, MAR Ferreira, D Bender, O Delaneau, J Marchini, and JF Zagury. Bias Characterization in Probabilistic Genotype Data and Improved Signal Detection with Multiple Imputation. *PLOS Genetics*, 12(6):e1006091, jun 2016.

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[37] Brenda L. Powell, Steven Wiltshire, Gillian Arscott, Pamela A. McCaskie, Joseph Hung, Brendan M. McQuillan, Peter L. Thompson, Kim W. Carter, Lyle J. Palmer, and John P. Beilby. Association of PARL rs3732581 genetic variant with insulin levels, metabolic syndrome and coronary artery disease. *Human Genetics*, 124(3):263–270, oct 2008.

[38] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. 2016.

[39] Charles Savona-Ventura and Stephanie Savona-Ventura. The inheritance of obesity. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 29(3):300–308, 2015.

[40] M. Sałaga, A. Mokrowiecka, P.K. Zakrzewski, A. Cygankiewicz, E. Leishman, M. Sobczak, H. Zatorski, E. Małecka-Panas, R. Kordek, M. Storr, W.M. Krajewska, H.B. Bradshaw, and J. Fichna. Experimental colitis in mice is attenuated by changes in the levels of endocannabinoid metabolites induced by selective inhibition of fatty acid amide hydrolase (FAAH). *Journal of Crohn's and Colitis*, 8(9), 2014.

23

[41] C R Scott, S H Clark, C C Teng, and K R Swedberg. Clinical and cellular studies of sarcosinemia. *The Journal of pediatrics*, 77(5):805–11, nov 1970.

[42] Bertrand Servin, Matthew Stephens, P Kraft, P Pharoah, SJ Chanock, D Albanes, LN Kolonel, N Li, M Stephens, P Scheet, M Stephens, L Almasy, J Blangero, EI George, RE McCulloch, JD Nott, PJ Green, X Yang, TR Belin, WJ Boscardin, M Stephens, N Smith, P Donnelly, M Stephens, P Scheet, M Lynch, B Walsh, H Jeffreys, AE Raftery, DJ Lunn, JC Whittaker, N Best, RE Kass, AE Raftery, I Good, G Hellenthal, M Stephens, CS Carlson, MA Eberle, MJ Rieder, Q Yi, L Kruglyak, JM Chapman, JD Cooper, JA Todd, DG Clayton, ME Weale, C Depondt, SJ Macdonald, A Smith, PS Lai, SK Tate, C Depondt, SM Sisodiya, GL Cavalleri, S Schorge, J Marchini, D Cutler, N Patterson, M Stephens, E Eskin, EL Heinzen, W Yoon, SK Tate, A Sen, NW Wood, S Zöllner, JK Pritchard, N Soranzo, GL Cavalleri, ME Weale, NW Wood, C Depondt, J Marchini, S Myers, G McVean, P Donnelly, I Pe'er, PIW De Bakker, J Maller, K Jones, MD Altshuler, B Servin, M Stephens, DL Nicolae, JY Dai, I Ruczinski, M LeBlanc, C Kooperberg, MJ Sillanpaa, M Bhattacharjee, AP Morris, J Cohen, A Pertsemlidis, IK Kotowski, R Graham, and CK Garcia. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS Genetics*, 3(7):e114, 2007.

[43] Shutong Shen, Lichan Tao, Xiuzhi Wang, Xiangqing Kong, and Xinli Li. Common Variants for Heart Failure. *Current genomics*, 16(2):82–7, apr 2015.

[44] So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, Klaudia Walter, Cristina Menni, Lu Chen, Louella Vasquez, Ana M Valdes, Craig L Hyde, Vicky Wang, Daniel Ziemek, Phoebe Roberts, Li Xi, Elin Grundberg, Melanie Waldenberger, J Brent Richards, Robert P Mohney, Michael V Milburn, Sally L John, Jeff Trimmer, Fabian J Theis, John P Overington, Karsten Suhre, M Julia Brosnan, Christian Gieger, Gabi Kastenmüller, Tim D Spector, Nicole Soranzo, and Nicole Soranzo. An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–550, may 2014.

[45] S. M. Smith, J. F. Gregory, G. H. Zeisel, C. R. Gibson, T. H. Mader, J. Kinchen, P. Ueland, R. Ploutz-Snyder, M. Heer, and S. R. Zwart. Risk of Visual Impairment and Intracranial Hypertension After Space Flight: Evaluation of the Role of Polymorphism of Enzymes Involved in One-Carbon Metabolism. 2016.

[46] G. K. Smyth. limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer-Verlag, New York, 2005.

[47] O Tange. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine*, 36(1):42–47, feb 2011.

[48] Vitor Hugo Teixeira, Robert Olaso, Marie-Laure Martin-Magniette, Sandra Lasbleiz, Laurent Jacq, Catarina Resende Oliveira, Pascal Hilliquin, Ivo Gut, François Cornelis, and Elisabeth Petit-Teixeira. Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. *PloS one*, 4(8):e6803, aug 2009.

[49] Omer Weissbrod, Christoph Lippert, Dan Geiger, and David Heckerman. Accurate liability estimation improves power in ascertained case-control studies. *Nature Methods*, 12(4):332–334, feb 2015.

[50] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, jan 2014.

[51] Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, Jin Chen, Martin L Buchkovich, Samia Mora, Jacques S Beckmann, Jennifer L Bragg-Gresham, Hsing-Yi Chang, Ayşe Demirkan, Heleen M Den Hertog, Ron Do, Louise A Donnelly, Georg B Ehret, Tõnu Esko, Mary F Feitosa, Teresa Ferreira, Krista Fischer, Pierre Fontanillas, Ross M Fraser, Daniel F Freitag, Deepti Gurdasani, Kauko Heikkilä, Elina Hyppönen, Aaron Isaacs, Anne U Jackson, Åsa Johansson, Toby Johnson, Marika Kaakinen, Johannes Kettunen, Marcus E Kleber, Xiaohui Li, Jian'an Luan, Leo-Pekka Lyytikäinen, Patrik K E Magnusson, Massimo Mangino, Evelin Mihailov, May E Montasser, Martina Müller-Nurasyid, Ilja M Nolte, Jeffrey R O'Connell, Cameron D Palmer, Markus Perola, Ann-Kristin Petersen, Serena Sanna, Richa Saxena, Susan K Service, Sonia Shah, Dmitry Shungin, Carlo Sidore, Ci Song, Rona J Strawbridge, Ida Surakka, Toshiko Tanaka, Tanya M Teslovich, Gudmar Thorleifsson, Evita G Van den Herik, Benjamin F Voight, Kelly A Volcik, Lindsay L Waite, Andrew Wong, Ying Wu, Weihua Zhang, Devin Absher, Gershim Asiki, Inês Barroso, Latonya F Been, Jennifer L Bolton, Lori L Bonnycastle, Paolo Brambilla, Mary S Burnett, Giancarlo Cesana, Maria Dimitriou, Alex S F Doney, Angela Döring, Paul Elliott, Stephen E Epstein, Gudmundur Ingi Eyjolfsson, Bruna Gigante, Mark O Goodarzi, Harald Grallert, Martha L Gravito, Christopher J Groves, Göran Hallmans, Anna-Liisa Hartikainen, Caroline Hayward, Dena Hernandez, Andrew A Hicks, Hilma Holm, Yi-Jen Hung, Thomas Illig, Michelle R Jones, Pontiano Kaleebu, John J P Kastelein, Kay-Tee Khaw, Eric Kim, Norman Klopp, Pirjo Komulainen, Meena Kumari, Claudia Langenberg, Terho Lehtimäki, Shih-Yi Lin, Jaana Lindström, Ruth J F Loos, François Mach, Wendy L McArdle, Christa Meisinger, Braxton D Mitchell, Gabrielle Müller, Ramaiah Nagaraja, Narisu Narisu, Tuomo V M Nieminen, Rebecca N Nsubuga, Isleifur Olafsson, Ken K Ong, Aarno Palotie, Theodore Papamarkou, Cristina Pomilla, Anneli Pouta, Daniel J Rader, Muredach P Reilly, Paul M Ridker, Fernando Rivadeneira, Igor Rudan, Aimo Ruokonen, Nilesh Samani, Hubert Scharnagl, Janet Seeley, Kaisa Silander, Alena Stancáková, Kathleen Stirrups, Amy J Swift, Laurence Tiret, Andre G Uitterlinden, L Joost van Pelt, Sailaja Vedantam, Nicholas Wainwright, Cisca Wijmenga, Sarah H Wild, Gonneke Willemsen, Tom Wilsgaard, James F Wilson, Elizabeth H Young, Jing Hua Zhao, Linda S Adair, Dominique Arveiler, Themistocles L Assimes, Stefania Bandinelli, Franklyn Bennett, Murielle Bochud, Bernhard O Boehm, Dorret I Boomsma, Ingrid B Borecki, Stefan R Bornstein, Pascal Bovet, Michel Burnier, Harry Campbell, Aravinda Chakravarti, John C Chambers, Yii-Der Ida Chen, Francis S Collins, Richard S Cooper, John Danesh, George Dedoussis, Ulf de Faire, Alan B Feranil, Jean Ferrières, Luigi Ferrucci, Nelson B Freimer, Christian Gieger, Leif C Groop, Vilmundur Gudnason, Ulf Gyllensten, Anders Hamsten, Tamara B Harris, Aroon Hingorani, Joel N Hirschhorn, Albert Hofman, G Kees Hovingh, Chao Agnes Hsiung, Steve E Humphries, Steven C Hunt, Kristian Hveem, Carlos Iribarren, Marjo-Riitta Järvelin, Antti Jula, Mika Kähönen, Jaakko Kaprio, Antero Kesäniemi, Mika Kivimaki, Jaspal S Kooner, Peter J Koudstaal, Ronald M Krauss, Diana Kuh, Johanna Kuusisto, Kirsten O Kyvik, Markku Laakso, Timo A Lakka, Lars Lind, Cecilia M Lindgren, Nicholas G Martin, Winfried März, Mark I McCarthy, Colin A McKenzie, Pierre Meneton, Andres Metspalu, Leena Moilanen, Andrew D Morris, Patricia B Munroe, Inger Njølstad, Nancy L Pedersen, Chris Power, Peter P Pramstaller, Jackie F Price, Bruce M Psaty, Thomas Quert-

ermous, Rainer Rauramaa, Danish Saleheen, Veikko Salomaa, Dharambir K Sanghera, Jouko Saramies, Peter E H Schwarz, Wayne H-H Sheu, Alan R Shuldiner, Agneta Siegbahn, Tim D Spector, Kari Stefansson, David P Strachan, Bamidele O Tayo, Elena Tremoli, Jaakko Tuomile-hto, Matti Uusitupa, Cornelia M van Duijn, Peter Vollenweider, Lars Wallentin, Nicholas J Wareham, John B Whitfield, Bruce H R Wolffenbuttel, Jose M Ordovas, Eric Boerwinkle, Colin N A Palmer, Unnur Thorsteinsdottir, Daniel I Chasman, Jerome I Rotter, Paul W Franks, Samuli Ripatti, L Adrienne Cupples, Manjinder S Sandhu, Stephen S Rich, Michael Boehnke, Panos Deloukas, Sekar Kathiresan, Karen L Mohlke, Erik Ingelsson, and Gonçalo R Abecasis. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11):1274–1283, oct 2013.

[52] Feng Wu, Themistocles Dassopoulos, Leslie Cope, Anirban Maitra, Steven R. Brant, Mary L. Harris, Theodore M. Bayless, Giovanni Parmigiani, and Shukti Chakravarti. Genome-wide gene expression differences in Crohn's disease and ulcerative colitis from endoscopic pinch biopsies: Insights into distinctive pathogenesis. *Inflammatory Bowel Diseases*, 13(7):807–821, jul 2007.

[53] Jian Yang, S Hong Lee, Michael E Goddard, Peter M Visscher, L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, T.A. Manolio, T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, et Al., B. Maher, E.E. Eichler, J. Flint, G. Gibson, A. Kong, S.M. Leal, J.H. Moore, J.H. Nadeau, J. Yang, B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, et Al., G. Gibson, B.J. Hayes, P.M. Visscher, M.E. Goddard, I. Strandén, D.J. Garrick, P.M. VanRaden, H.D. Patterson, R. Thompson, A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, D. Reich, D. Falush, M. Stephens, J.K. Pritchard, A.R. Gilmour, R. Thompson, B.R. Cullis, C.R. Henderson, L.E. Kruuk, M.E. Goddard, N.R. Wray, K. Verbyla, P.M. Visscher, G. de Los Campos, D. Gianola, D.B. Allison, S.M. Purcell, N.R. Wray, J.L. Stone, P.M. Visscher, M.C. O'Donovan, P.F. Sullivan, P. Sklar, International Schizophrenia Consortium, H. Lango Allen, K. Estrada, G. Lettre, S.I. Berndt, M.N. Weedon, F. Rivadeneira, C.J. Willer, A.U. Jackson, S. Vedantam, S. Raychaudhuri, et Al., J.W. Kent, T.D. Dyer, J. Blangero, M. Lynch, B. Walsh, D.S. Falconer, E.R. Dempster, I.M. Lerner, A. Robertson, I.M. Lerner, S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, P.C. Sham, C.D. Campbell, E.L. Ogburn, K.L. Lunetta, H.N. Lyon, M.L. Freedman, L.C. Groop, D. Altshuler, K.G. Ardlie, J.N. Hirschhorn, L.R. Cardon, L.J. Palmer, R.R. Hudson, L. Liang, S. Zöllner, G.R. Abecasis, C.J. Hoggart, M. Chadeau-Hyam, T.G. Clark, R. Lampariello, J.C. Whittaker, M. De Iorio, D.J. Balding, C.C. Spencer, Z. Su, P. Donnelly, J. Marchini, Y. Li, G.R. Abecasis, Y.S. Aulchenko, S. Ripke, A. Isaacs, C.M. van Duijn, Wellcome Trust Case Control Consortium, A.R. Gilmour, B.J. Gogel, B.R. Cullis, and R. Thompson. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76–82, jan 2011.

[54] Minglan Yang, Lei Ye, Bokai Wang, Jie Gao, Ruixin Liu, Jie Hong, Weiqing Wang, Weiqiong Gu, and Guang Ning. Decreased miR-146 expression in peripheral blood mononuclear cells is correlated with ongoing islet autoimmunity in type 1 diabetes patients 1miR-146. *Journal of Diabetes*, 7(2):158–165, mar 2015.

[55] Scheid S Yang X and Lottaz C. OrderedList: Similarities of Ordered Gene Lists.

[56] Jin Zheng, Yun Li, Gonçalo R. Abecasis, and Paul Scheet. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genetic Epidemiology*, 35(2):102–110, feb 2011.